University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2018

Automatically creating a spatially referenced corpus of landscape perception

Chesnokova, Olga ; Purves, Ross S

Abstract: Spatially referenced thematically relevant corpora are an important first step in analyzing a wide variety of phenomena. Here, we describe and evaluate a workflow which extracts descriptions containing first person perception of landscape, and associates these with polygon geometries used in characterizing landscapes.

DOI: https://doi.org/10.1145/3281354.3281356

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-162271
Conference or Workshop Item
Published Version

Originally published at:
Chesnokova, Olga; Purves, Ross S (2018). Automatically creating a spatially referenced corpus of landscape perception. In: GIR 2018 : 12th Workshop on Geographic Information Retrieval at ACM SIGSPATIAL 2018, Seattle, 6 November 2018, ACM Digital Library.
DOI: https://doi.org/10.1145/3281354.3281356

Automatically creating a spatially referenced corpus of landscape perception

Olga Chesnokova
University of Zurich
Zurich, Switzerland
olga.chesnokova@geo.uzh.ch

Ross S. Purves
University of Zurich
Zurich, Switzerland
ross.purves@geo.uzh.ch

ABSTRACT

Spatially referenced thematically relevant corpora are an important first step in analyzing a wide variety of phenomena. Here, we describe and evaluate a workflow which extracts descriptions containing first person perception of landscape, and associates these with polygon geometries used in characterizing landscapes.

CCS CONCEPTS

• **Information systems** → *Content analysis and feature selection*; *Document collection models*; Data extraction and integration;

KEYWORDS

toponym, landscape, corpus, GIR

ACM Reference Format:
Olga Chesnokova and Ross S. Purves. 2018. Automatically creating a spatially referenced corpus of landscape perception. In *12th Workshop on Geographic Information Retrieval (GIR'18), November 6, 2018, Seattle, WA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3281354.3281356

1 INTRODUCTION

A wide range of application areas require the collection of thematically relevant, spatially located documents for further analysis [5]. In many cases documents should be linked to existing geographies – for example polygon boundary data associated with political, ecological or cultural phenomena. Here we describe the development of a workflow to associate textual descriptions with areas of distinctive landscape character used in the planning process.

Our aim is to extract descriptions containing landscape perception, since landscapes are, in part, defined by how they are perceived by individuals experiencing them [3]. Current approaches to assessing landscape perception focus on expert knowledge and empirical work in the landscape (for example through interviews). However, such information is also present in web documents (e.g. travel blogs) which describe first person experiences in a landscape. The following text characterizing a region is found in a report produced using traditional methods: "[...] Predominantly a very tranquil landscape due to the openness and perceived naturalness [...]; and minimal sources of artificial noise [...]" [6, p. 35, section 5.0] and can immediately be seen to have many parallels with the following descriptions: "[...] each time we have been up on the tops have hardly seen anybody - today was no exception."[1], "We haven't seen a road or heard a car for about 7hrs."[2].

In this paper we develop a prototype workflow to extract and spatially reference such descriptions automatically. For a small test area in the English Lake District, we assess both how common such descriptions are for ten existing areas of distinctive landscape character and our ability to associate descriptions with individual areas.

2 METHODS

To create our spatially referenced, thematically relevant corpus of texts we first retrieved potentially relevant documents with a web-crawler. After coarse filtering of irrelevant documents using terms found within URLs, we classified the remaining documents according to whether they contained first person perception of landscapes. Finally, we assigned these documents to official polygons of distinctive landscape character.

We initially selected 10 neighboring areas of distinctive character (from a total of 71) in the English Lake District delineated using traditional methods [6]. Fig. 1 shows an extract from the larger area, illustrating the complexity of the polygon borders. On average, these regions have an area of $27km^2$, and each is referred to using a set of characteristic toponyms, typically names of hills, settlements, lakes or valleys (e.g. "Broom, Ling, Kirk Fells"). These toponyms were used as seed terms to build an initial document corpus using the BootCaT toolkit [1]. A first coarse filtering was performed using URLs to remove irrelevant documents (e.g. holiday rentals, hotels, local government information). After filtering, the remaining documents were annotated with respect to three classes:

- First person perception of landscape, e.g. "A thankfully short unpleasant section through conifers, no sound, no vegetation and hardly any light"[3]
- Landscape descriptions which do not describe individual experiences, for example descriptions of guided walks, e.g. "Routes starting from Skiddaw Forrest in the east are also quieter, giving the walker a sense of being in the wilderness."[4]
- not relevant, e.g. cottage descriptions, official parish information, weather forecasts, etc.

The distinction between the first two classes is important for our task, since we aim to generate a corpus of personal feelings and individual experiences, rather than descriptions presenting an

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
GIR'18, November 6, 2018, Seattle, WA, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6034-0/18/11…$15.00
https://doi.org/10.1145/3281354.3281356

[1] http://www.masarnenramblers.com/lords-seat-barf-broom-fell–graystones.html
[2] http://www.walkingforum.co.uk/index.php?topic=29872.0
[3] https://www.andrewswalks.co.uk/lordsseatgroup.html
[4] http://english-lake-district.info/skiddaw/skiddaw.html

idealized view as often used in promoting or describing activities more generally. After annotation, we used a random forest classifier to classify thematically relevant documents (see section 3).

The final step in our workflow was to spatially link individual descriptions to areas of distinctive landscape character in the Lake District. To do so we first performed toponym recognition using an existing Ordnance Survey gazetteer limited to our study area [2]. Gazetteer lookup was carried out using unigrams, bigrams and trigrams, fuzzy matching with Levenshtein distance, and specific rules for common generic terms used in compound nouns and where capitalization is often inconsistent (e.g. Derwentwater/ Derwent Water/ Derwent water). Toponym resolution was carried out using DBScan [4]. Since first person perception descriptions of landscape mostly describe walks and other recreational activities, we judged this simple approach adequate to both disambiguate and remove outliers (such as distant peaks seen but not visited). The final toponym set contained a set of point locations, each associated with a frequency. To allocate descriptions to areas of distinctive landscape character we firstly created three classes of toponym frequency based on Jenks natural breaks. The area containing the most frequently used toponym (and the highest toponym count in the case of ties) was then associated with this description. Since descriptions may be associated with more than one polygon, we then applied a simple region growing approach based on topological adjacency.

3 RESULTS

We used BootCaT to query for our 10 areas with 15 toponyms, and retrieved a total of 641 documents, after filtering, from a total possible of 1500 (since BootCaT restricts us to 100 documents per query). We found an average of 42.7 documents per area (median 42) of which on average 6.8 were first person perception of landscape (median 2). Interestingly, toponym type had a very strong influence on the number of documents retrieved, with the 5 toponyms containing the most frequent relevant first person descriptions (average 17.4, median 18 descriptions) all referring to hills.

Our random forest classification of first person perception of landscape used the following features: presence of selected personal pronouns; most frequent unigrams and 50 terms with the highest document frequency per class and part-of-speech category. The overall precision of the random forest classifier, trained on half of the data, was 0.84.

For 10 random texts we evaluated the quality of our toponym recognition using our gazetteer look-up method. Average precision was 0.86 and recall 0.79. We did not evaluate the quality of our toponym resolution directly, since we were interested in the efficacy of our approach in allocating documents to polygons. To evaluate we therefore manually allocated 20 randomly chosen descriptions to one or more areas of distinctive landscape character. By comparing the annotated data with our algorithmic solution, we measured overall precision as 0.86 and recall as 0.70. Furthermore, for all 20 descriptions at least one area of distinctive landscape character was correctly identified.

4 CONCLUDING DISCUSSION

The biggest limitation with respect to retrieving relevant documents was our use of BootCaT, which returned a relative small

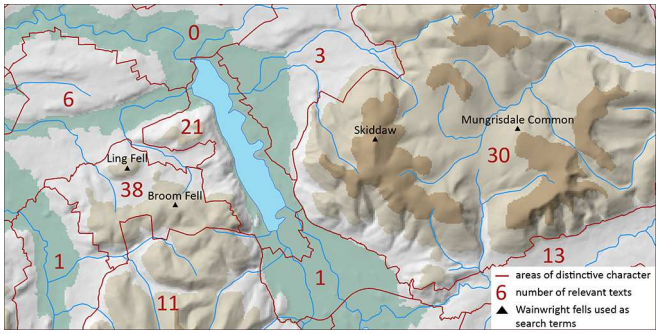

Figure 1: Number of documents associated with areas of distinctive landscape character

number of documents. Our classification and spatial referencing steps both had good precision, as illustrated by the distribution of documents associated with our 10 seed polygons. For 3 polygons BootCaT retrieved no relevant documents as a result of strong filtering of content, semantic ambiguity (a toponym which is also a common surname) and feature type (village names are less often associated with first person perception). Precision for both classification and georeferencing was high (0.84 and 0.86 respectively) implying that we were able to find and locate relevant descriptions successfully. However, these descriptions were strongly associated with the feature type used in our search. In our case, this is a result of individuals "collecting" Wainwrights, hills described and named in a series of books (c.f. Fig. 1). The influence of these lists was further demonstrated by the collection of correctly georeferenced documents to the south of our seed polygons. These are the result of our use of the toponym Kirk Fell, which is geographically ambiguous, but in Wainwright's list refers to a peak where we found a cluster of documents. This points to the importance of understanding external context influencing the production of documents relevant to a thematic corpus. We now plan to use our workflow, with a less restrictive web-crawler, to extract documents for the whole study area. More generally, we demonstrate how using relatively simple, but context-dependent rules, we can create high quality thematic and spatially referenced corpora by seeding initial search with fine-grained and specific toponym types.

REFERENCES

[1] Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC* 4 (2004), 1313–1316.

[2] Davide Buscaldi and Bernardo Magnini. 2010. Grounding Toponyms in an Italian Local News Corpus. In *Proceedings of the 6th Workshop on Geographic Information Retrieval - GIR '10.* https://doi.org/10.1145/1722080.1722099

[3] Council of Europe. 2000. European Landscape Convention. *Report and Convention Florence* ETS No. 17, 176 (2000), 8. http://conventions.coe.int/Treaty/en/Treaties/Html/176.htm

[4] Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, and Mauro Gaio. 2014. Geocoding for texts with fine-grain toponyms : an experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.*

[5] Jan Oliver Wallgrün, Morteza Karimzadeh, Alan M. MacEachren, and Scott Pezanowski. 2018. GeoCorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science* 32, 1 (2018), 1–29. https://doi.org/10.1080/13658816.2017.1368523

[6] Dominic Watkins. 2008. *Lake District National Park Landscape Character Assessment and Guidelines.* Technical Report. 466 pages.